Il Progetto di digitalizzazione Google books presso le biblioteche della Sapienza, Università di Roma

Adriana Magarotto - Maura Quaquarelli - Mattia Vallania
Sistema bibliotecario Sapienza – SBS

Introduzione ed elementi di contesto

Le biblioteche di ricerca hanno come missione quella di costruire nel tempo collezioni di documenti per rispondere ai bisogni della loro comunità di utenti. È necessario inoltre garantire l'accesso perpetuo a queste collezioni e facilitare il più possibile la diffusione e l'accesso alla memoria storica e al sapere custodito nelle biblioteche.

Le recenti imprese di digitalizzazione, i problemi di finanziamento e di spazio per la crescita delle raccolte a stampa hanno creato nuove opportunità e modificato la concezione stessa di biblioteca di ricerca su larga scala. Al fine di perseguire questa trasformazione, ormai in atto in tutte le biblioteche del mondo, Sapienza ha scelto di aderire al progetto Google Books, occasione unica per l'avvio di una digitalizzazione massiva. Altri progetti di digitalizzazione sono stati realizzati precedentemente da parte di singole strutture dell'Ateneo o nell'ambito del Progetto collaborativo ProDigi (2008-2009) e hanno consentito di creare un primo archivio di testi digitalizzati ma soprattutto di diffondere nell'Ateneo le conoscenze e le esperienze di base, oggi in parte riutilizzabili per la gestione di questo nuovo progetto molto più impegnativo e finalizzato alla realizzazione e incremento di Sapienza Digital Library[1].

A seguito dell'accordo tra Sapienza Università di Roma e il Ministero per i beni e le attività culturali - Mibac, siglato nel luglio del 2011, la Sapienza si è impegnata a mettere a disposizione per la digitalizzazione una parte del patrimonio librario di sua proprietà. Il progetto è subito entrato nella fase operativa consentendo l'invio dei primi volumi da novembre 2012. Si prevede di poter scansionare i primi 30.000 volumi nell'arco di un anno, iniziando dalla collezione di libri più antichi, con data di pubblicazione fino al 1872. La data del 1872 è il limite temporale imposto dalla legislazione internazionale sui diritti degli autori e degli editori.

La fase iniziale del progetto, considerato il poco tempo disponibile per le fasi preparatorie e il carattere sperimentale sia per l'Ateneo che per Google, è stata avviata con dieci biblioteche, a cui se ne sono aggiunte altre venti dall'inizio del 2013.

[1] http://sapienzadigitallibrary.uniroma.it/.

L'accordo con Google prevede la massima riservatezza sulle soluzioni tecniche di loro proprietà. In questa sede, pertanto, ci limiteremo a illustrare le problematiche affrontate, le scelte operate e il risultato raggiunto unicamente dal punto di vista dell'Ateneo, corredando il resoconto tecnico con alcuni elementi di riflessione emersi durante lo svolgimento dell'attività.

Caratteristiche organizzative

Volendo schematizzare l'impegno gestionale richiesto, si può riassumere il flusso di lavoro in tre operazioni fondamentali:
– Individuare le opere idonee alla digitalizzazione in base alla data di pubblicazione, al formato e allo stato di conservazione.
– Selezionare un item per ogni manifestazione, corredandolo di un identificativo univoco e di un file di metadati in formato MARC21/XML.
– Posizionare i volumi sui carrelli forniti da Google e produrre degli elenchi ordinati dei volumi contenuti in essi comprendenti anche la stima del valore commerciale, a fini assicurativi.

Il reale contesto operativo ha da subito evidenziato come queste operazioni - concettualmente semplici - presentassero dei problemi inediti rispetto all'amministrazione della pur complessa struttura bibliotecaria di Sapienza e come questi fossero in buona parte connaturati allo stato attuale dell'automazione dei servizi bibliotecari dell'Ateneo: è la soluzione a queste difficoltà, a nostro parere, a rappresentare il contributo originale meritevole di analisi e lo stimolo a una riflessione sulle future politiche di gestione del Sistema Bibliotecario Sapienza (SBS)[2].

Approcci operativi

Le problematiche affrontate riguardano quindi il formato dei dati bibliografici, la movimentazione dei volumi e l'organizzazione del flusso di lavoro in una struttura distribuita tra SBS e le biblioteche partecipanti.
I principali requisiti tecnici richiesti da Google per procedere alla digitalizzazione sono i seguenti:
– Le collezioni della Sapienza devono essere considerate come un'unica raccolta libraria pur essendo, come noto, suddivise in molte biblioteche distinte, dal punto di vista fisico, della titolarità dei diversi centri di spesa e dal punto di vista organizzativo.
– Il materiale deve essere corredato di metadati in formato MARC21/XML.

[2] Il Sistema Bibliotecario Sapienza, istituito a novembre 2011, è operativo dalla primavera del 2012, e, secondo le indicazioni del Regolamento organizzativo emanato con Decreto n. 4461 del 15/12/2011, art. 2: «ha lo scopo di assicurare la conservazione, lo sviluppo, la valorizzazione e la gestione integrata dell'intero patrimonio bibliografico e documentario della Sapienza, nonché l'accesso alle risorse informative on line in funzione delle esigenze della ricerca, della didattica e dell'Amministrazione».

– I volumi debbono essere dotati di un identificativo univoco a livello d'Ateneo, sottoforma di barcode tramite il quale viene gestita l'identificazione dei volumi durante il processo di scansione e la successiva associazione delle immagini digitali ai relativi metadati, al fine della loro visualizzazione nell'interfaccia pubblica di Google books.

Risulterà evidente come per rispondere a tali requisiti sia sufficiente riaggregare, all'uopo, i metadati descrittivi ed amministrativo-gestionali presenti nella base-dati di un moderno ILS.
Sapienza gestisce le procedure delle proprie biblioteche per mezzo di Sebina Open Library (SOL) un *integrated library system* prodotto da Data Management; il software, conforme al protocollo SbnMarc per il colloquio con l'Indice SBN, gestisce le principali funzioni biblioteconomiche (catalogazione, acquisti, prestito, gestione utenti, etc.) sia per il back che front-office, nonché l'export dei dati nei formati UNIMARC e MARC21.
Viste le peculiari esigenze del progetto, e i tempi brevi richiesti per l'organizzazione, si è valutato che l'applicazione, pur avendo molta flessibilità nelle modalità di export, non potesse, a meno di importanti interventi evolutivi, consentire rapidamente una corretta elaborazione dei dati, essendo necessarie operazioni ben diverse dalle funzioni per cui il software è stato sviluppato.
Per brevità, ricorderemo solo le principali criticità incontrate, ovvero (i) l'individuazione esatta dei testi da sottoporre a digitalizzazione in base alla data di pubblicazione e (ii) la necessità di selezionare un solo esemplare di una manifestazione da inviare per la digitalizzazione.
La struttura del dato bibliografico nel formato SBN prevede l'indicazione di una o due date di pubblicazione in campi codificati ma, in caso di data di pubblicazione incerta, considera i campi data come facoltativi. Pertanto, quando non sia presente alcuna data sul documento e quando non si ritenga utile indicare «le date estreme entro le quali si presume sia stata edita la pubblicazione»[3], i campi data non vengono valorizzati, con l'ovvio limite di non poter interrogare direttamente il catalogo, tramite l'ILS, in tutti quei casi in cui la data sia incerta e non siano stati valorizzati i campi facoltativi. Limitarsi a selezionare i testi per la digitalizzazione tramite SOL, pertanto, avrebbe escluso tutte le opere di datazione incerta, ma sicuramente rispondenti ai requisiti richiesti (si pensi, per esempio, tendendo a mente il 1872 come anno limite per l'idoneità, a un'opera con data F senza i campi data 1 e data 2 valorizzati, ma che riporti in descrizione la data 17...).
Quanto all'unicità degli esemplari da inviare in digitalizzazione, invece, va considerato come la complessa storia delle collezioni e la ramificazione delle biblioteche

[3] Tale norma è attualmente oggetto di revisione (cfr. Guida alla catalogazione in SBN materiale moderno, draft ottobre 2013).

dell'Ateneo fa sì che siano spesso presenti più items della medesima manifestazione in diverse biblioteche; di questi, tuttavia, non è possibile conoscere a priori lo stato di conservazione e le effettive condizioni di disponibilità così da indicare con precisione quale esemplare richiedere e a quale biblioteca. Ciò considerato si è subito reso evidente come fosse necessario un procedimento che a posteriori (ovvero una volta ricevuto il giudizio di idoneità dalle biblioteche) selezionasse una sola copia per le successive elaborazioni: come normale, anche tale singolarissima procedura non è presente nell'ILS.
Lo sviluppo di nuovi moduli all'interno del gestionale sarebbe stato senz'altro possibile, ma è apparso da subito un'opzione poco auspicabile per diverse ragioni. Intervenire su un prodotto così complesso richiede dei tempi di sviluppo e di test che non erano compatibili con i tempi molto ristretti dettati dalle esigenze di progetto (si consideri che con la soluzione alternativa illustrata di seguito le prime biblioteche hanno cominciato a comunicare la disponibilità dei volumi dopo appena una settimana dall'avvio del progetto); la necessità di adattare rapidamente lo strumento di lavoro in relazione alle esigenze operative (molto mutevoli nelle prime fasi) richiedeva di avere alcune componenti del sistema implementabili rapidamente anche da personale con competenze tecniche non specialistiche; un ruolo non marginale, infine, è stato giocato dalla valutazione dei costi di sviluppo.
Al di là dei singoli aspetti, emerge la constatazione di quanto siano cambiate e si siano evolute le esigenze informative di "un organismo che cresce" come le biblioteche accademiche e, di conseguenza, sia sempre più necessaria una maggiore flessibilità che consenta di utilizzare strumenti nuovi e diversi a fronte di nuove esigenze.
Le esigenze ora ricordate hanno portato allo sviluppo di un insieme di tools informatici affidato ad un team composto da personale di SBS delle biblioteche della Sapienza e dai tecnici del Cineca.
Tale ambiente operativo, descritto in dettaglio di seguito, gestisce tutti gli aspetti delle lavorazioni: l'interfaccia per il bibliotecario, che formula il giudizio di idoneità per i volumi della propria biblioteca, i controlli di unicità per l'esemplare, la produzione dei metadati e della documentazione cartacea di accompagnamento.
Al cuore del sistema, come ovvio, una base dati relazionale necessaria a tracciare gli stati di lavorazione e a effettuare i necessari controlli di coerenza sui dati.
Disporre delle informazioni bibliografiche su una base dati indipendente ne ha consentito un'analisi avanzata: si consideri, a titolo esemplificativo, l'analisi condotta sul già citato problema di estrazione delle date di pubblicazione. In mancanza di informazioni nei campi Data 1 e Data 2 non rimane che la data nell'area della pubblicazione della descrizione ISBD (nel suo recepimento SBN). Dal punto di vista tecnico, l'area della pubblicazione si presenta come una stringa alfanumerica dalla formulazione piuttosto complessa all'interno della quale è possibile isolare l'indicazione della data mediante un parsing basato sulle regole di punteggiatura previste dallo standard per delimitare gli elementi all'interno delle aree della de-

scrizione bibliografica. Tale processo di analisi della stringa è normalmente eseguito per la traduzione dei dati nei principali formati di interscambio di dati bibliografici, riportando la data indicata in descrizione in un sottocampo del tracciato previsto dal formato in uso (UNIMARC 210$d, $h per la data di stampa e MARC 21 260$c). Tuttavia le descrizioni bibliografiche presenti nel nostro catalogo, in parte frutto di riversamenti o di campagne di catalogazione approssimative, spesso non rispettano la punteggiatura prescritta dallo standard e di conseguenza i dati frutto della traduzione MARC sono poco utilizzabili.
Grazie a un'analisi separata delle descrizioni ISBD nella base dati di appoggio, si è potuto affiancare all'esame dei sottocampi MARC una seconda interrogazione formulando una procedura di matching della data (basata sull'uso di espressioni regolari), che tenga conto degli errori di catalogazione più frequentemente riscontrati, aumentando così di circa il 10% i risultati dell'interrogazione sul solo campo MARC[4].
L'impatto con questo tipo di problemi se da un lato fornisce l'ennesima prova empirica della necessità per i bibliotecari di ibridare le proprie competenze in materie e procedimenti solo apparentemente lontani dalla propria professionalità, dall'altro spinge a un'evoluzione dei rapporti con i partner commerciali nella direzione di un *business model* orientato non più solo alla vendita del software chiavi in mano, ma anche all'offerta di supporto, formazione e personalizzazione.

Flusso di lavoro

1 Attribuzione del giudizio di idoneità secondo le specifiche di progetto.

Per mezzo delle procedure appena citate SBS seleziona, per ogni biblioteca partecipante, i volumi idonei in base alla data di pubblicazione; quindi predispone i record estratti in un'interfaccia mediante la quale le biblioteche possono completare la valutazione di idoneità in base a tre elementi:
– Effettive condizioni di disponibilità del volume e stato di conservazione
– Idoneità relativa ai parametri fisici (dimensioni) richiesti per la digitalizzazione
– Stima del valore del volume, calcolata secondo un algoritmo sviluppato dalla comunità bibliotecaria di Sapienza.

L'interfaccia di inserimento dati è costituita da un foglio di calcolo on-line *Google spreadsheet*, uno per biblioteca.
I file vengono messi a disposizione tramite un sito ad accesso riservato, che contiene anche tutte le informazioni relative alle procedure di lavoro da adottare.

2. "Deduplicazione" dei record e generazione dell'identificativo univoco (*barcode*)

I dati elaborati dai bibliotecari vengono quindi importati nel DBMS di gestione (MySql) per le lavorazioni successive. La procedura che popola il database, realiz-

[4] Analisi condotta su un campione limitato di dati.

zata in Php dal Cineca, sfrutta le *Google spreadsheet API* per automatizzare la procedura di import.
Una volta completato l'import dei dati dichiarati idonei dalle varie biblioteche e prima della generazione dell'identificativo univoco viene operata, da un'apposita procedura, la cosiddetta "deduplicazione" di eventuali esemplari multipli di una stessa edizione, selezionando il primo esemplare dichiarato idoneo; le altre copie vengono automaticamente scartate.
Viene quindi stampato il *barcode* per tutti i record idonei.

Barcode – specifiche tecniche
La stringa, codificata nel *barcode* secondo lo standard code 39, si compone di venti caratteri più la cifra di controllo. Tutti i *barcode* riportano RMS come caratteri iniziali, sigla del sistema bibliotecario; seguono le due cifre che identificano la biblioteca di provenienza e quindici caratteri con l'inventario assegnato da questa al volume. Tutti i caratteri vuoti, così come richiesto da Google per facilitare la ricerca a partire dal *barcode*, sono stati sostituiti da una cifra normalmente non utilizzata all'interno degli inventari, il $.
La scelta di generare un identificativo univoco composto da elementi significativi (e non per esempio da sequenze casuali univoche) è stata dettata dalla volontà di disporre di un ID che stabilisse di per se il legame con l'oggetto reale.

Barcode: RMS2LOLD000009467$$$F	
RMS: codice polo	**2L**: codice Biblioteca di Filosofia
OLD000009467$$$: inventario OLD 9467	**F**: cifra di controllo

Figura 1. *Esempio stringa codificata nel barcode*

3. Registrazione del prestito e spedizione

I *barcode* vengono quindi inviati alle biblioteche per essere posti nei volumi, prima del loro posizionamento sui carrelli.
La generazione delle liste di volumi che verranno inviati è a carico delle biblioteche: gli operatori, mentre sistemano fisicamente i volumi sui carrelli predisposti per il trasporto, registrano il prestito in SOL e, sulla base di queste registrazioni, viene generato l'elenco dei volumi secondo l'ordine in cui sono disposti. Mediante la sequenza temporale delle registrazioni dei prestiti (timestamp), è infatti possibile generare un elenco, che corrisponde esattamente all'ordine dei volumi sul carrello, come richiesto dalle specifiche del progetto.
Nel database di gestione viene quindi importato l'elenco di tutti i volumi registrati in prestito; tali informazioni, oltre a tracciare la movimentazione del patrimonio, servono a produrre in modo automatico la documentazione cartacea d'accompagno da inserire in ogni carrello e la lista complessiva dei volumi inviati.

Nel catalogo in linea compare l'informazione che il volume non è disponibile per il tempo necessario (3 mesi).

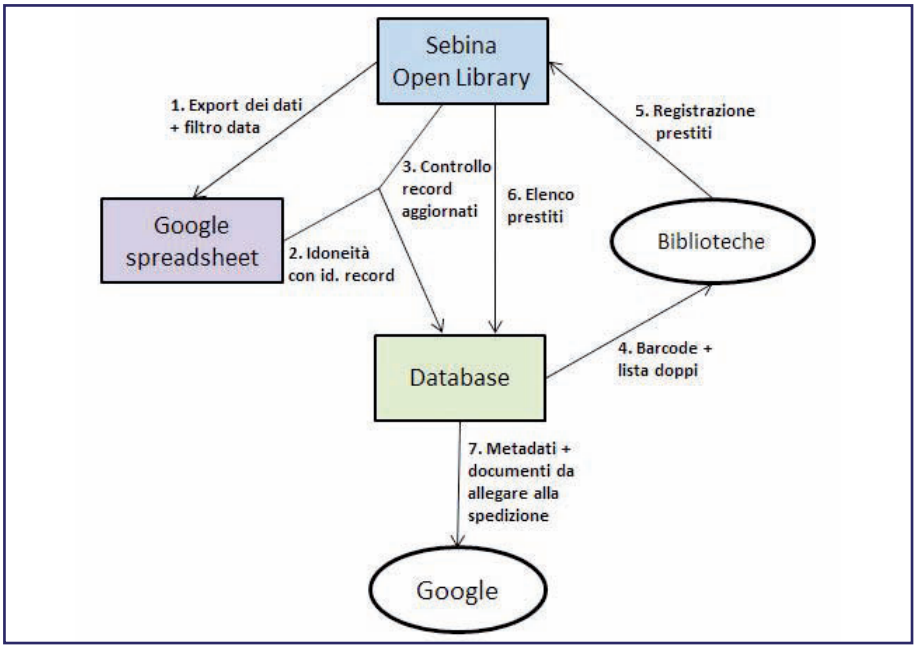

Figura 2. *Schema del flusso di lavoro*

4. Produzione metadati

L'export dei dati bibliografici in MARC21 ottenuto dal software SOL viene convertito in MARC 21/XML[5] secondo lo schema standard; per rispondere alle richieste di Google, tuttavia, sono necessarie alcune manipolazioni sia sui metadati amministrativi che descrittivi.

Le specifiche di progetto, infatti, prevedono un file di metadati MARC21/XML per ogni pezzo fisico inviato in digitalizzazione, completo di tutte le informazioni necessarie all'identificazione dell'opera, richiedendo che tutte le informazioni relative al titolo siano presenti nel campo 245.

Il manuale SBN, invece, prevede la catalogazione multilivello (fino a tre livelli) per tutte le pubblicazioni in più volumi (con l'unica eccezione delle opere in più volumi indivisibili). Questo ha comportato delle modifiche - piuttosto laboriose nel caso di catalogazioni a tre livelli - per adeguare i files alle specifiche: se infatti nel caso di due soli livelli è stato sufficiente operare sul singolo file, spostando il contenuto del campo 773 (relativo alla relazione cosiddetta padre/figlio) in 245, nel caso di catalogazioni a tre livelli si è reso necessario sviluppare una procedura che

[5] http://www.loc.gov/standards/marcxml//.

interroga il catalogo risalendo i legami gerarchici per ricostruire l'intera relazione.
Il campo 245 è stato quindi opportunamente modificato utilizzando i sottocampi $n e $p per gestire il contenuto dei legami gerarchici.
Si è inoltre provveduto a completare la notizia con le informazioni riguardanti la responsabilità intellettuale dell'opera; le regole SBN, infatti, prevedono che in alcuni casi il legame con l'autore sia stabilito solo sul record padre e quindi, per restituire l'integrità dell'informazione in un singolo file, è necessario riportare tali legami nel record figlio.
Per la parte amministrativo-gestionale, invece, viene aggiunta nel campo 955 la stringa dell'identificativo univoco stampato sottoforma di *barcode* e incluso nei volumi.

```
<record>
   <leader>00729nam  22001697i 4500</leader>
   <controlfield tag="001">UBO0326879</controlfield>
   <controlfield tag="008">040802s1844    it ||||  |    |||||||ITAod</controlfield>
   <datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">IT-SOL</subfield>
   </datafield>
   <datafield tag="041" ind1="0" ind2=" ">
      <subfield code="a">ITA</subfield>
   </datafield>
   <datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Romagnosi, Gian Domenico</subfield>
   </datafield>
   <datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Opere del professore G. D. Romagnosi</subfield>
      <subfield code="p">13: Dell'indole e dei fattori dell'incivilimento</subfield>
      <subfield code="c">del professore G. D. Romagnosi</subfield>
   </datafield>
   <datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">Firenze</subfield>
      <subfield code="b">nella Stamperia Piatti</subfield>
      <subfield code="c">1844</subfield>
   </datafield>
   <datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">XI, 314 p.</subfield>
      <subfield code="c">21 cm.</subfield>
   </datafield>
   <datafield tag="774" ind1=" " ind2="0">
      <subfield code="d">Firenze : nella stamperia Piatti</subfield>
      <subfield code="t">Opere del professore G. D. Romagnosi</subfield>
      <subfield code="w">RMS912557</subfield>
      <subfield code="8">13</subfield>
   </datafield>
   <datafield tag="850" ind1=" " ind2=" ">
      <subfield code="a">RMS2L</subfield>
   </datafield>
   <datafield tag="955" ind1=" " ind2=" ">
      <subfield code="a">BIBLIOTECA DI FILOSOFIA</subfield>
      <subfield code="z">RMS2LOLD000009467$$$F</subfield>
   </datafield>
</record>
```

Figura 3. *Esempio record bibliografico in MARC21/XML*

Risultati e sviluppi futuri

L'incremento delle biblioteche coinvolte nel progetto e il conseguente aumento delle notizie trattate ha fatto emergere notevoli difficoltà nel tentativo di mantenere allineati i tre sistemi con cui si lavora, SOL, i Google spreadsheet e il database di gestione. Le procedure di controllo verso il catalogo vengono effettuate ogniqualvolta si debba procedere a produrre dei documenti, siano questi i barcode o le liste di volumi caricate nei carrelli, ma questo implica che qualsiasi operazione venga rallentata dagli allineamenti che è necessario fare nel caso i controlli non vadano a buon fine. Ad oggi, infatti, non è previsto un update automatico dei dati, pertanto è sempre l'operatore che deve ripristinare manualmente la coerenza di questi.

Escludendo la possibilità di gestire l'intero progetto tramite SOL, per le ragioni illustrate in precedenza, riteniamo comunque opportuno proseguire lo sviluppo del sistema in un'ottica di riduzione delle origini di dati eliminando quindi l'utilizzo degli spreadsheet (che contengono, si ricorda, un export filtrato dal catalogo) e permettendo alle biblioteche di lavorare su dati aggiornati in tempo reale. Con la collaborazione del Cineca, stiamo lavorando a un'interfaccia che permetta alle biblioteche di visualizzare i dati presenti su SOL in un'interfaccia Web ad accesso riservato e di inserire lì tutte le informazioni utili al progetto.

Altro aspetto ancora da ingegnerizzare è la gestione del rientro dei volumi nelle biblioteche dopo la scansione, che vada ad integrare il semplice "rientro dal prestito" con le informazioni sullo stato della lavorazione messe a disposizione da Google in un'interfaccia Web, che consente di sapere se il volume è stato digitalizzato e, in caso negativo, le ragioni che lo hanno escluso dalla procedura.

Attualmente tali dati vengono comunicati alle biblioteche mediante tabelle in Excel, ma l'intenzione è quella di importare le informazioni nel database di gestione per comunicarle automaticamente alla biblioteche e completare così il tracciamento del percorso di ogni volume.

Rimane, infine, da sviluppare la gestione delle immagini digitali che restituisce Google, che devono essere convertite e inserite nel nostro repository, Sapienza Digital Library, e integrati nel catalogo SOL.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2013.